Data Engineering in Azure

Course Duration: 20 hours

Course Syllabus:

1. Introduction to Database and Programming
2. Introduction to Data Engineering and Azure Storage
3. Data Lake and Data Storage Architecture
4. Data Warehousing with Azure Synapse Analytics
5. ETL and Data Integration using Azure Data Factory
6. Real-Time Data Processing and Streaming
7. Advanced Data Engineering with Azure Databricks
8. Security and Monitoring
9. Capstone Project

Tools & Technologies Covered:

1. SQL
2. Python
3. Azure Blob Storage & Data Lake Gen2
4. Azure SQL Database
5. Azure Synapse Analytics
6. Azure Data Factory (ADF)
7. Azure Databricks
8. Azure Event Hubs
9. Azure Stream Analytics
10. Azure Key Vault
11. Azure Monitor

01 Introduction to Database and Programming

- SQL - Basics
- Python – Basics, PySpark, Pandas, NumPy, Requests, Azure SDK for python

02 Introduction to Data Engineering and Azure Storage

- **Data Engineering Basics:** Data ingestion, transformation, storage, and analysis.
- **Azure tools:** Azure Data Factory, Synapse Analytics, and Databricks.
- **Storage Services Overview:**
  - Introduction to Azure Storage (Blob Storage, Azure Data Lake Gen2).
  - Importance of scalable and secure data storage in the cloud.
- **Hands-on Lab:**
  - Create a storage account using Azure Portal.
  - Explore creating storage accounts and databases via Azure CLI.

03 Data Lake and Data Storage Architecture

- Introduction to Data Lake and Azure Data Lake Storage Gen2
- Differences between Azure Data Lake Storage and Blob Storage.
- **Hands-on Labs:**
  - Create Azure Data Lake Storage through the Portal.
  - Upload and organize files (e.g., CSV, JSON).

04 Data Warehousing with Azure Synapse Analytics

- **Synapse Analytics Introduction:**
  - Unified analytics service for big data and data warehousing.
- **Compute Options:**
  - Serverless SQL: Pay-as-you-use query execution.
  - Dedicated SQL Pools: Optimized for large-scale, high-performance workloads.
- **External Tables:**
  - Query external files stored in Azure Blob or Data Lake.
- **Hands-on Labs:**
  - Create Synapse Analytics Workspace.
  - Implement Dedicated SQL Pool.
  - Use notebooks for data processing in Spark Pools.

05 ETL and Data Integration using Azure Data Factory

- **ADF Overview:**
  - Extract, Transform, and Load (ETL) pipelines to move and transform data.
- **Parameterization in ADF:**
  - Automating workflows with parameterized pipelines and datasets.
- **Data Cleansing:**
  - Handling null values, schema drift, and conditional splits.
- **Hands-on Labs:**
  - Convert CSV to Parquet.
  - Load JSON into SQL.
  - Set up monitoring and alerts in ADF.

06 Real-Time Data Processing and Streaming

- **Azure Event Hubs:** Streaming platform for real-time data ingestion.
- **Azure Stream Analytics**: Real-time event processing and output to Power BI.
- **Hands-on Labs:**
  - Stream data from Event Hubs and process using Azure Stream Analytics.
  - Create Power BI visualizations from real-time data.

07 Advanced Data Engineering with Azure Databricks

- Introduction to Azure Databricks: Scalable Spark-based analytics service.
- **Delta Lake:** Transactional storage layer for reliability and performance.
- **Data Processing with Spark:** Handling diverse formats (JSON, Parquet).
- **Hands-on Labs:**
  - Create Databricks cluster.
  - Perform data cleaning and visualization in Databricks notebooks.
  - Implement a Delta Lake pipeline.

08 Security and Monitoring

- **Data Security in Azure:**
  - Azure Key Vault, encryption, and access control lists (ACL).
  - Row-level and column-level security in Synapse.
- **Monitoring:** Using Azure Monitor to track pipeline performance.
- **Hands-on Labs:**
  - Configure Key Vault.
  - Set up Synapse security (e.g., row-level security).

09 Capstone Project

T**itle:** Sales Reporting Pipeline in Azure

Objective: Build an end-to-end data pipeline to collect, process, and analyze sales data using Azure services.

- **Data Ingestion:** Use Azure Data Factory (ADF) to ingest sales data from Blob Storage and customer data from Azure SQL Database.
- **Data Storage:** Store raw data in Azure Data Lake Storage Gen2.
- **Data Transformation:** Process and aggregate data with Azure Databricks (PySpark).
- **Data Loading:** Load transformed data into Azure Synapse Analytics for querying.
- **Reporting:** Connect Power BI to Synapse for visualizing sales trends and product performance.